

## 基于分布式计算平台的海量时空数据可视查询推理框架

笔记本： 2015.08.01 Spark Pipeline

创建时间： 10/18/2015 11:36 AM

更新时间： 10/19/2015 7:27 AM

作者： yuxinm

---

## 动机

- 基于可视分析方法的查询（Query）+推理（Reasoning）：强调在海量数据场景下，普通可视分析系统的处理能力低下。
- 基于Spark平台的时空数据分析：强调使用分布式平台处理海量时空数据。
- 基于工作流的查询或分析流程构建：在Spark平台下缺少交互式构建工作流的工具和方法

## 贡献

针对海量时空数据的可视查询推理框架

- 通过工作流方式构建时空数据的查询。
- 通过可视化增强对查询结果的认知。
- 提出在Spark框架下的in-situ探索方法，用户可以随时看到查询或分析过程的运行状态。
- 融合采样：用户可以从一个局部样本或全局采样样本出发进行渐进式查询。

## 数据

基站数据

- 轨迹抽取
- 结合同一时段的社交数据（？）
- 新的数据组织形式，以优化Spark平台的查询处理。（？）

## 分析框架

- 查询模块及分析模块构建：使用工作流方式
- 数据采样方法：结合可视化方法的采样策略（？）
- 对Spark上的计算过程进行可视化：通过中间结果输出
- 对Spark原生的查询语句（Spark QL）监控：监控查询进度

## 可视分析方法

总括：查询+推理的证据链

- 两个基本元素：查询+推理
- 设计类似知识图谱的结构，用于外化用户的查询推理网